2012-05-31

Validating the Veracity of User Data Collected and Disclosed by Ad Networks

Can Tatar
Worcester Polytechnic Institute

Validating the Veracity of User Data Collected and Disclosed by Ad Networks

by

Can Tatar

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

May 2012

APPROVED:

_____

Professor Craig E. Wills, Advisor and Department Head

_____

Professor Robert E. Kinicki, Reader

Abstract

The use of behavioral targeting practices provides ad networks with the opportunity to tailor ads to the individual characteristics of users. As privacy concerns over behavioral targeting have been growing lately, an increasing number of ad networks offer *ad preferences managers* (APMs) that show collected and/or inferred information about users. The focus of our study is to investigate the accuracy and completeness of the information contained in such APMs. On the basis of our experimental results, we propose a structured methodology for APM validation. We also assess how third parties render ads based on users' browsing behavior. Our findings reveal cases in which even sensitive information is leaked as part of an HTTP header and is used to serve ads on multiple sites. The third parties examined in this study include an intent-focused data exchange (BlueKai) and a social network (Facebook) along with the ad networks owned by AOL, Google, and Yahoo!.

Acknowledgments

I am deeply grateful to my advisor, Craig E. Wills, for his guidance and continued support. It was a privilege to work with him and to benefit from his vast experience.

I would also like to thank my reader, Robert E. Kinicki, for his thorough review of this thesis.

Over the past two years, I was fortunate to have been funded as a TA at WPI. I thank the faculty in the CS Department, especially the professors I TAed for and the members of the PEDS Research Group.

Finally, I would like to thank my family for always being there for me and for supporting me in everything I do.

Table of Contents

List of Tables

Table Page

List of Figures

Figure Page

Chapter 1: Introduction

The use of behavioral targeting practices provides online advertising networks with the opportunity to tailor advertisements to the individual characteristics of users. Third-party HTTP cookies are the most common means of tracking users' browsing behavior [25, 41].

An increasing number of ad networks offer *ad preferences managers* (APMs) that show collected and/or inferred information about users. After Google launched a behavioral advertising program in March 2009 [18], privacy advocates praised its decision to provide users with access to their APMs and with an "opt-out" mechanism, but they also indicated Google needed to do more to let people know that their behavior was being monitored. The majority of American adults do not want ads to be rendered based on their interests and they believe that the law should obligate advertisers "to immediately delete information about their internet activity" [42].

A typical APM created by an aggregator consists of the following elements: areas of interest, IP-based/gathered geographic information (which may include zip code) as well as demographic variables such as age, gender, and income. Technical attributes (e.g., operating system, browser, screen resolution, and color depth) used to display ads appropriately for the user's environment are also provided by some companies like Yahoo! [47]. Users can edit and/or remove most pieces of information in their APMs, primarily their interest categories. Moreover, most of the major ad networks allow users to opt out of ads targeted to their online behavior.

In this work, we propose a methodology for testing the accuracy and completeness of the information in APMs made available to users by ad networks. We also assess how third parties render ads based on users' browsing behavior. Our findings reveal cases in which even sensitive information is leaked as part of an HTTP header and is used to serve ads on multiple sites. The third parties examined in this study include an intent-focused data exchange (BlueKai) and a social network (Facebook) along with the ad networks owned by AOL, Google, and Yahoo!.

The rest of this thesis is structured as follows: Chapter 2 covers the background information. The methodology we used to study third parties is described in Chapter 3. The test results and the fundamental aspects of the Facebook test are given in Chapters 4 and 5 respectively. Finally, Chapter 6 presents the conclusions and future directions.

Chapter 2: Background

The rise of computational advertising has triggered systematic research on related systems. An experimental study [16] proposed a measurement methodology for online advertising networks. It defined new metrics that are resistant to significant levels of noise present in ad distribution. Wang et al. [45] developed an ad auditing methodology and demonstrated that it is capable of effectively monitoring ad networks on a large scale. A client-side method for detecting and classifying third-party trackers based on how they manipulate browser state was introduced in [36]. Another relevant study [38] described how online ad exchanges work and found that the complexity of these systems provides criminals with an opportunity to gain revenue by writing malware that mimics legitimate user activities.

As public concern over behavioral targeting has been growing lately, several studies, such as [12, 17, 34, 35, 41], have focused on privacy-preserving online advertising. A novel way to handle third-party cookies that allows users to have

control over information available to aggregators was introduced in [12]. Guha et al. [17] presented a system which maintains profiles locally on the users' computers rather than in the cloud. A browser extension was proposed in [41] that enables the targeting algorithm to run in the browser. Riederer et al. [35] proposed a mechanism called 'transactional' privacy, in which users decide what information about themselves is put up for *sale* while receiving compensation for it, and third parties purchase *access* to exploit this information for ad targeting. Another recent study [34] addressed the problem of running auctions that leverage the information in the user profile for ad ranking while keeping the profile private.

As demonstrated in [27, 28], leakage of behavioral data to third-party servers has been increasing dramatically in the past few years. Just as importantly, an additional study [26] showed that a third party can link the users' browsing behavior to personal information and identifiers using data mined from online social networks (OSNs) that employ its and/or its affiliated third parties' services. In a more recent study, Krishnamurthy et al. [24] examined over 100 popular non-OSN Web sites and found leakage on 75% of these sites. Korolova [23] proposed a new class of attacks that exploit the microtargeting capabilities of Facebook's advertising system in order to violate user privacy.

In attempting to understand the sensitivity of information contained in behavioral profiles, it is noteworthy to consider that privacy researcher Latanya Sweeney's study [39] revealed that 87% of the U.S. population can be uniquely identified solely by their 5-digit zip code, birth date, and gender. Furthermore, Narayanan and Shmatikov [31] emphasized the possibility of re-identification without personally identifiable information (PII). They noted that "any information that distinguishes one person from another can be used for re-identifying anonymous data."

Current policy and technology research on third-party Web tracking is discussed in [30] based on the results from a new dynamic Web measurement platform, FourthParty. Numerous instances of non-compliance with behavioral advertising notice and choice requirements were identified in [22]. Wills [46] characterized the necessary conditions for private user information made available to a first-party site to be leaked to a third party and provided specific instances of where leakage occurs. He also showed how this leakage can be prevented through a number of actions available to end users as well as first-party sites.

According to [20], users' privacy preferences regarding sharing their locations with advertisers are complicated, and offering advanced privacy settings is helpful in mitigating their concerns about location-based advertising. As reported in [21],

users are typically willing to trade off ease-of-use against higher levels of control over their personal information and are thus more comfortable with an explicit profiling system.

2.1 Summary

The rise of computational advertising has triggered systematic research on related systems. There is considerable ongoing effort to develop structured measurement methodologies for online advertising networks. As public concern over behavioral targeting has been growing lately, several studies have focused on privacy-preserving online advertising. Also, a number of researchers studied leakage to third-party aggregators, the threat of re-identification of anonymous data, and users' privacy preferences.

Chapter 3: Methodology

The third parties offering ad preferences managers (APMs) to end users without the need for an account include 33Across [1], AOL [3], Bizo [4], BlueKai [6], eXelate [10], Google [14], interclick [19], Lotame [29], TARGUSinfo [40], Videology [44], and Yahoo! [47]. We opted to examine the ad networks owned by Web giants Google, AOL, and Yahoo! and online data exchange BlueKai due to its distinct characteristics. BlueKai does not serve ads—it operates an online platform for intent-focused data exchange [5, 32]. Names of the APMs of the third parties examined in this study are listed in Table 1.

Table 1: Names of the ad preferences managers of the third parties examined in this study

Third party	**Name of APM**
AOL	AdVisibility—My Advertising Preferences
BlueKai	BlueKai Registry
Google	Ads Preferences Manager
Yahoo!	Ad Interest Manager

We compiled a list of 15 first-party sites for each third party. We used a browser extension that automatically visits over 1000 popular sites and lists the third parties these sites utilize [9]. We identified the sites to be included in the test sets mostly using this list. We also included sites that fall into sensitive categories (health and sexual orientation) or that offer a profile page on which sensitive information can be provided.

We ran daily sessions for a ten-day period for all third parties. A session comprises visiting 15 Web sites successively and performing pre-defined actions on these sites (such as reading technology news on `nytimes.com` or searching for information on skin cancer on `medhelp.org`). The same controlled browser (either IE, Chrome, or Opera) was used for an entire test—i.e., 10 daily sessions. Only before the first session, we deleted all cookies and history in the test browser. Whenever we needed to log on to a Web site, we did so and never logged out of it. During the sessions, we captured the ads served by the relevant ad server and checked the APM after visiting each site. We used the information Web traffic analysis companies Alexa [2], Compete [7], and Quantcast [33] provide about the first-party sites to verify the information shown in the APM. Nevertheless, it is possible that an interest category is generated based on what is displayed on a Web site we visit even if that category does not pertain to the intended nature of

the site. We also recorded the HTTP traffic and inspected the HTTP headers using Fiddler [11], an HTTP(S) debugging proxy, to find out what is passed to the ad server in plaintext. However, it should be remembered that leakage can be deliberate or inadvertent—i.e., third parties may not be expecting all information they receive.

To classify ads, we used the following categories:

Expected types of advertising. The advertising practices that are expected to be observed are as follows:

- generic: No information pertinent to the user is used.
- location-based: IP-based geographic information is used.
- contextual: The ad is related to what is currently displayed on the page.
- behavioral—in APM: The advertised product/service directly relates to previous online behavior, and a relevant category, at a minimum, is shown in the APM.
- based on profile on that page: A piece of information that is in the profile on that page is used.

Unexpected types of advertising. The advertising practices that are not expected to be observed are as follows:

- behavioral—not in APM: The advertised product/service directly relates to previous online behavior, although a relevant category is *not* shown in the APM.
- based on previous profile: A piece of information that is available in a profile on a previously-visited page is used.
- based on sensitive behavior: The ad is related to past sensitive behavior (visiting a gay site, searching for a disease on a health site, etc.).

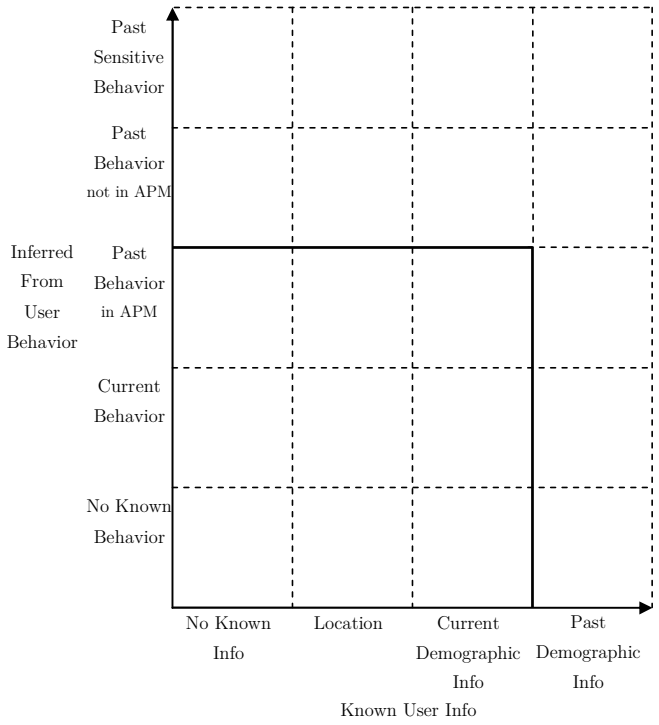

Figure 1: Two-dimensional classification of online advertising practices (inner square: expected results, remaining boxes: unexpected results)

We developed a two-dimensional classification to illustrate the advertising practices we observed during our tests (see Figure 1). The inner square is used to show expected results, and the remaining boxes are used to show unexpected results.

The subsequent subsections list the first-party sites (and their categories based on the information obtained from Alexa [2], Compete [7], and Quantcast [33]) used to study each third party.

3.1 Google

Table 2: First-party sites used to study Google

Site	**Category**
accuweather.com	News, Weather
bloomberg.com	News, Business
cbsnews.com	News
gaylife.about.com	Gay People
linkedin.com	Professional Networking
macmillandictionary.com	Dictionaries, English
medhelp.org	Health
metrolyrics.com	Music, Lyrics
nytimes.com	News, Newspapers
pandora.com	Radio
snagajob.com	Employment, Job Search
tmz.com	News, Entertainment, Celebrities
toyota.com	Automotive, Toyota
tripadvisor.com	Travel
yelp.com	Consumer Opinions

The actions performed on the sites listed in Table 2 included searching for "Miami" (on `accuweather.com`, `tripadvisor.com`, and `yelp.com`) and for "skin cancer" (on `medhelp.org`).

3.1.1 Disabling third-party cookies and opting out

We performed an additional Google test by disabling third-party cookies and another one by opting out of customized Google Display Network ads to understand whether these mechanisms are truly useful for preventing behavioral targeting of ads.

After 10 sessions with the sites listed in Table 2, we enabled third-party cookies (opted in during the other test) and ran 3 extra sessions with 5 other sites that are listed in Table 3. This enabled us to see if we were targeted during these extra sessions based on our behavior in the first 10 sessions.

Table 3: First-party sites visited in the extra Google sessions

Site	**Category**
`ehow.com`	How-To Guide
`imdb.com`	Movies
`reference.com`	Information Reference
`target.com`	Retailers
`wunderground.com`[1]	News, Weather

[1] We searched for the weather forecast in New York on this site.

3.2 AOL

Table 4: First-party sites used to study AOL

Site	**Category**
autoblog.com	Automotive
autos.aol.com	Automotive
cars.com	Automotive
encyclopedia.com	Encyclopedias
engadget.com	News, Technology
huffingtonpost.com	News
latimes.com	News, Newspapers
match.com	Dating
music.aol.com	Music
shoutcast.com	Radio
slashcontrol.com†	News, Television († now huffingtonpost.com/tv)
spinner.com	Music
techcrunch.com	News, Technology
theboot.com	Music, Country/Bluegrass Music
tourtracker.com	Music, Tours, Concerts

The actions performed on the sites listed in Table 4 included searching for Toyota dealers in Miami (on cars.com) and for information on Miami (on encyclopedia.com). Our user was listed as a man interested in men on match.com.

3.3 Yahoo!

Table 5: First-party sites used to study Yahoo!

Site	**Category**
autos.yahoo.com	Automotive
espanol.yahoo.com	Portals, Spanish
health.yahoo.net	Health
hotwire.com	Travel
lyricsmode.com	Music, Lyrics
mercurynews.com	News, Newspapers
monster.com	Employment, Job Search
nissanusa.com	Automotive, Nissan
ozonebilliards.com	Shopping, Sports, Cue Sports
realestate.yahoo.com	Real Estate
sfgate.com	News, Newspapers
travel.yahoo.com	Travel
webmd.com	Health
wunderground.com	News, Weather
xfinity.comcast.net	Service Providers, Cable

The actions performed on the sites listed in Table 5 included searching for "skin cancer" (on webmd.com), for "diabetes" (on health.yahoo.net), and for "Miami" (on realestate.yahoo.com, travel.yahoo.com, hotwire.com, monster.com, and wunderground.com). Our user's location was listed as Seattle on monster.com.

3.4 BlueKai

Table 6: First-party sites used to study BlueKai

Site	**Category**
accuweather.com	News, Weather
azlyrics.com	Music, Lyrics
bankrate.com	Personal Finance
bmwusa.com	Automotive, BMW
cars.com	Automotive
cbsnews.com	News
cdkitchen.com	Cooking
cheaptickets.com	Travel
contactmusic.com	Entertainment, Music, Movies
gap.com	Shopping, Clothing
healthology.com	Health
miami.com	Guides, Miami
money.cnn.com	News, Business
style.com	Fashion
zdnet.com	News, Technology

The actions performed on the sites listed in Table 6 included searching for the weather forecast in Chicago (on `accuweather.com`), for flights from New York to Chicago (on `cheaptickets.com`), and for "skin cancer" (on `healthology.com`).

3.5 Test with Google properties after March 1, 2012

There were only minor differences between our generic methodology and the methodology we used for the Google test we performed after Google's recent privacy policy change [15]. In this test, we aimed to observe the potential use of information collected on Google properties to serve ads on non-Google sites. We therefore replaced three sites in the list given in Table 2 (`metrolyrics.com`, `snagajob.com`, `tmz.com`) with popular Google properties (Google+, Google Search, and YouTube).

We created a Google account, signed in, and did not sign out of it throughout the test. Note that we *unchecked* the box shown in Figure 2 when creating the account.

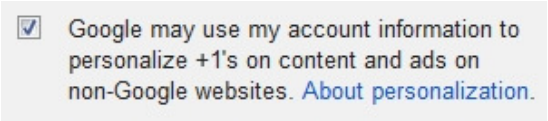

Figure 2: Personalization preference when creating a Google account

3.6 Summary

We opted to examine the ad networks owned by Web giants Google, AOL, and Yahoo! and online data exchange BlueKai due to its distinct characteristics. For all third parties, we ran daily sessions with 15 first-party sites for a ten-day period. During the sessions, we captured the ads served by the relevant ad server and checked the APM after visiting each site. We used the information Web traffic analyzers provide about the sites to verify the information shown in the APM. We also recorded the HTTP traffic and inspected the HTTP headers to find out what is passed to the ad server in plaintext. Finally, we identified expected and unexpected types of advertising and developed a two-dimensional classification to illustrate the advertising practices we observed during our tests.

Chapter 4: Results

4.1 Google

A sample snapshot of the Google Ads Preferences Manager from our tests is shown in Figure 3. The demographic information is inferred based on the sites visited [14]. However, the age range has changed several times throughout the tests, although we kept visiting the same set of sites listed in Table 2 and performing similar actions on these sites.

Figure 4 summarizes Google's advertising practices we were able to identify. We adopted the terminology "some sessions" to refer to multiple, but no more than half of the sessions, and "most sessions" to refer to more than half, but not all, of the sessions. Light gray shading represents some, gray represents most, and dark gray represents all sessions.

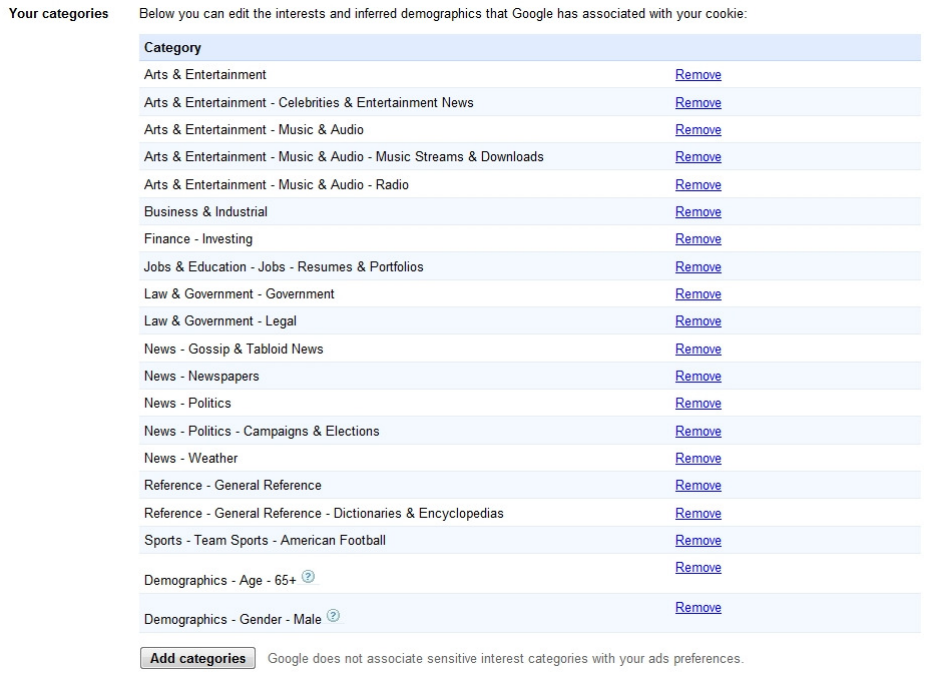

Figure 3: A sample snapshot of Google's Ads Preferences Manager

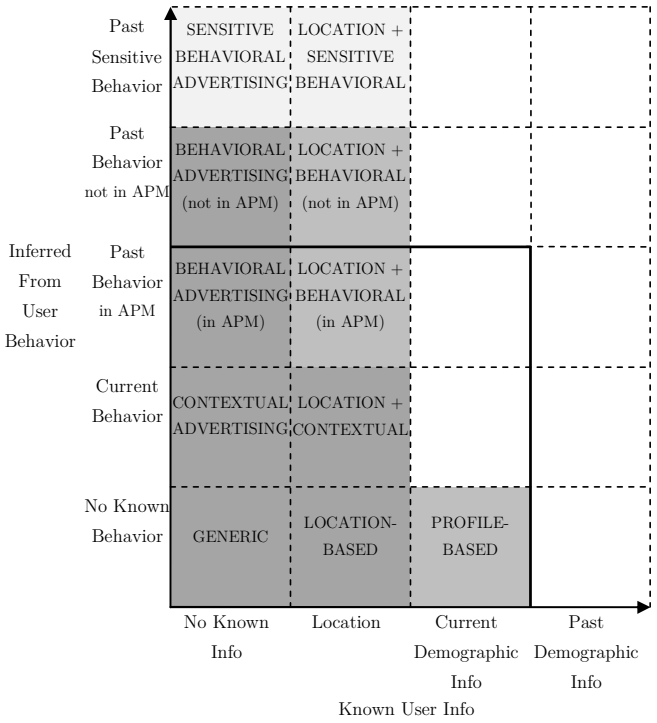

Figure 4: Classification of Google's advertising practices based on our test results (inner square: expected results, remaining boxes: unexpected results; light gray: observed in some sessions, gray: most sessions, dark gray: all sessions)

Note that, in [13], it is stated that "Google will not associate sensitive interest categories with your cookie (such as those based on race, religion, sexual orientation, health, or sensitive financial categories) and will not use these categories when showing you interest-based ads."

As an example, Figure 5 demonstrates leakage to DoubleClick[1] on LinkedIn.

```
GET
/adi/linkedin.dart/home;optout=false;lang=en;tile=2;sz=300x250;v=4;u=sjta
jT8Or6xLr51Oe6R4kkR7;mod=250;title=en;func=acct;coid=3881;ind=68;csize=h;
zip=60637;cntry=us;reg=14;sub=0;jpos=0;con=a;edu=18319;gy=2002;gdr=m;seg=
499;sjt=40;extra%3Dnull;s=0;ord=264085484? HTTP/1.1
Host: ad.doubleclick.net
Referer: http://www.linkedin.com/home
```

Figure 5: Leakage of personal information to DoubleClick from a LinkedIn profile

We now provide sample DoubleClick (`doubleclick.net`) ads for each advertising practice seen in Figure 4.

Generic advertising. No information pertinent to the user is used by the third party to serve ads that fall into this category. The ad shown in Figure 6 is an example of generic ads served by DoubleClick.

[1] DoubleClick, a subsidiary of Google, is the Internet's largest advertising company [37].

Figure 6: A generic ad served by DoubleClick on `tmz.com`

Location-based advertising. Figure 7 shows an example of ads served by DoubleClick based on IP-based geographic information.

Figure 7: A location-based ad served by DoubleClick on `accuweather.com`

Contextual advertising. Figure 8 presents an example of contextual ads served by DoubleClick.

Figure 8: A contextual ad served by DoubleClick on `gaylife.about.com`

Location-based+contextual advertising. Figure 9 shows an example of ads rendered based on the context of the site and geographic location of the IP address.

Figure 9: A location-based+contextual ad served by DoubleClick on medhelp.org

Behavioral advertising. Figure 10 provides an example of ads customized based on the past online behavior of our user.

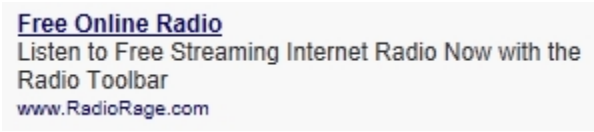

Figure 10: A behavioral ad served by DoubleClick on `nytimes.com` after listening to music on Pandora

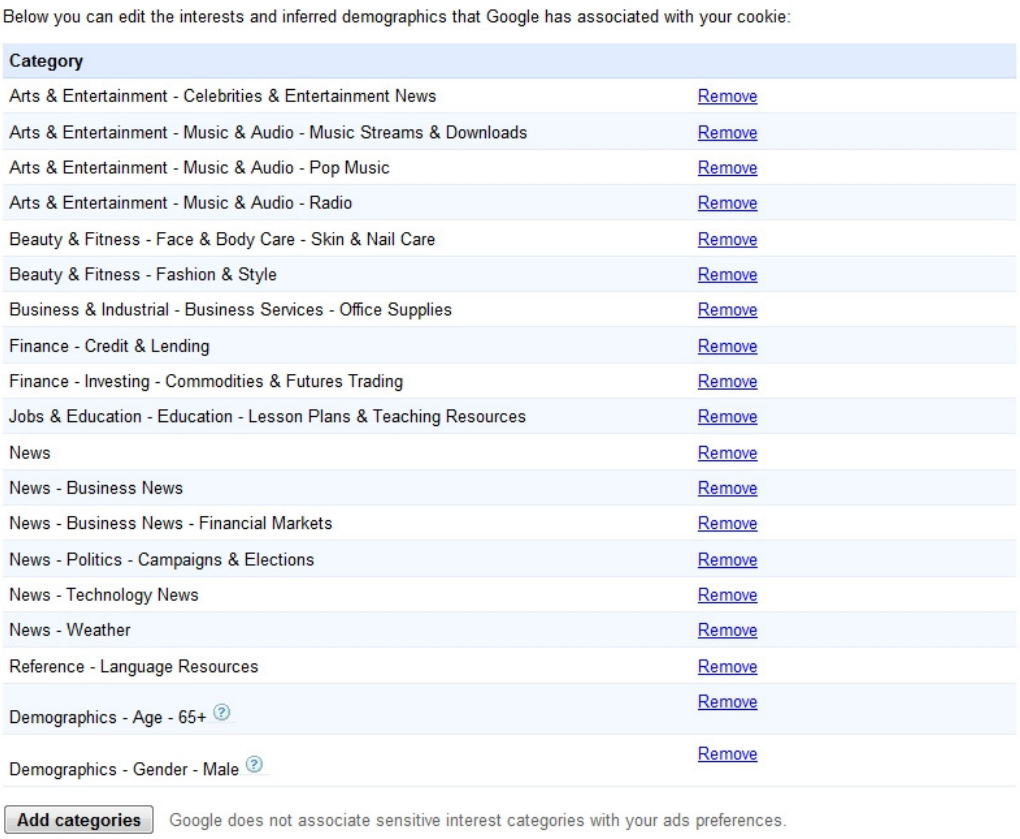

Figure 11: Google Ads Preferences Manager while visiting `nytimes.com` during a test session

Several types of behavioral advertising are indicated in Figure 4. Note that there were a couple of "Music & Audio" categories in the Google Ads Preferences Manager (see Figure 11) at the time we observed the ad shown in Figure 10.

In some cases, previous online behavior was combined with the current site's context to serve ads (see Figure 12).

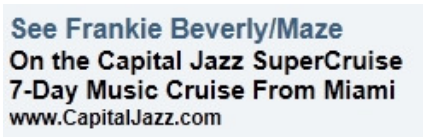

Figure 12: A contextual+behavioral ad served by DoubleClick while listening to smooth jazz on `pandora.com` after searches for "Miami" on `accuweather.com`, `tripadvisor.com`, and `yelp.com`

Figure 13 shows how DoubleClick learned about our search for the weather forecast in Miami on `accuweather.com`. Although Google's APM has the category "World Localities - North America - USA - Florida - Southern Florida - South Florida Metro - Miami-Dade," this was never shown.

```
GET /adj/accuwx.us.forecast/city-weather-
forecast;zip=33128;city=miami;state=fl;country=us;partner=accuweather;met
ro=mia;ctrav=1;strav=1;cuwx=7;fc1wx=16;fc1hi=85;fc1lo=75;fc2wx=17;fc2hi=9
0;fc2lo=75;fc3wx=17;fc3hi=90;fc3lo=75;ulang=tr;vabeachtemp=70;vabeachwx=4
;ixc=10101;pos=top;sz=980x30,728x90;tile=1;ord=691599837061021000?
HTTP/1.1
Host: ad.doubleclick.net
Referer: http://www.accuweather.com/us/fl/miami/33128/city-weather-
forecast.asp
```

Figure 13: Leakage to DoubleClick while performing a search on `accuweather.com`

Location-based+behavioral advertising. Figure 14 presents an example of ads rendered based on the past online behavior of our user and geographic location of the IP address.

Figure 14: A location-based+behavioral ad served by DoubleClick on `accuweather.com` after visiting a page about Tiger Woods (an American golfer) on TMZ Sports

It should be noted that, at the time we observed the ad shown in Figure 14, Google Ads Preferences Manager did *not* list any relevant interest categories (see Figure 15).

Figure 15: Google Ads Preferences Manager while visiting `accuweather.com` during a test session

Profile-based advertising. We listed the location as Chicago in our user's Pandora profile and confirmed that this information was leaked to DoubleClick (see Figure 16).

```
GET /adj/pand.default/prod.radio;index=1;interaction=station;fam=-
1;artist=G162;gcat=g111g461;genre=rock;ag=32;gnd=1;zip=60637;hours=0;comp
ed=0;exp=0;fb=0;dma=602;clean=0;msa=005;st=IL;co=17031;et=0;pin=0;aa=1;hi
sp=0;hhi=0;u=index*1!interaction*station!fam*-
1!artist*G162!gcat*g111g461!genre*rock!ag*32!gnd*1!zip*60637!hours*0!comp
ed*0!exp*0!fb*0!dma*602!clean*0!msa*005!st*IL!co*17031!et*0!pin*0!aa*1!hi
sp*0!hhi*0;tile=1;sz=2000x2;ord=1329614846952195567 HTTP/1.1
Host: ad.doubleclick.net
Referer: http://www.pandora.com/radioAdEmbed.html?cb=13296148469807485
```

Figure 16: Leakage of the zip code (along with other information) from a Chicago-based user's Pandora profile

Figure 17 shows an example of ads based on the information in a user profile—a Pandora profile in this case. Note that the ad was served on Pandora as well.

Figure 17: A profile-based ad served by DoubleClick on `pandora.com`

Sensitive behavioral advertising. We also observed Google ads related to our user's past sensitive behavior as indicated in Figure 4. As part of our tests, we

searched for "skin cancer" on `medhelp.org`, and this search query was passed to DoubleClick as shown in Figure 18.

```
GET
/adi/medhelp.search/search;area=search;logged_in=no;ss=6;action=index;env
=production;tile=1;sz=728x90;site=medhelp;dcopt=ist;position=leader;dc_re
f=http%3A%2F%2Fwww.medhelp.org%2Fsearch%3Futf8%3D%2526%2523x2713%253B%26q
uery%3Dskin%2Bcancer%26camp%3Dtop_nav_search;ord=6972645583994780?
HTTP/1.1
Host: ad.doubleclick.net
Referer:
http://www.medhelp.org/search?utf8=%26%23x2713%3B&query=skin+cancer&camp=
top_nav_search
```

Figure 18: Leakage of a sensitive search query to DoubleClick from a medical Web site (`medhelp.org`)

We then observed cancer-related ads on multiple Web sites. An example of those ads is shown in Figure 19.

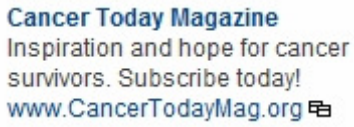

Figure 19: A sensitive behavioral ad served by DoubleClick on the "Tech and Science News" page of `cbsnews.com`

One of the sites we visited was `gaylife.about.com`, and we observed many ads targeted at people who are interested in men on multiple other sites. An example of such ads is shown in Figure 20. None of these ads included the word "gay," but the inferred gender of our user was male.

Figure 20: A *suspected* sensitive behavioral ad served by DoubleClick on `macmillandictionary.com`

Location-based+sensitive behavioral advertising. During our tests, we received Google ads not just from DoubleClick, but from other ad servers of Google, including `2mdn.net`, `googleadservices.com`, and `googlesyndication.com`. Some of these ads included sensitive behavioral information. For instance, the ad shown in Figure 21 was served by Google Ad Services on `pandora.com` after searches for "skin cancer" on `medhelp.org`. Note that this search term was actually leaked to DoubleClick as shown in Figure 18. Google Ad Services may have learnt our interest in skin cancer from the context of the "Search Results" page on `medhelp.org`, where it also served ads related to skin cancer, or it may have received this information from DoubleClick.

Figure 21: A sensitive behavioral ad served by Google Ad Services on `pandora.com`

Given that Tufts Medical Center is located in Boston, Massachusetts, and we ran our tests in Massachusetts, the ad shown in Figure 21 seems to utilize our geographic location as well.

The Google test results are summarized in Table 7.

Table 7: Google test results

Site / Search term	**Category**	**Google Ads Preferences Manager**	**Behavioral ads shown**
accuweather.com	News, Weather	News - Weather	✓
bloomberg.com	News, Business	Business & Industrial	✓
cbsnews.com	News	News	✓
gaylife.about.com	Gay People	—	*Suspected*
linkedin.com	Professional Networking	—	—
macmillandictionary.com	Dictionaries, English	Dictionaries & Encyclopedias	✓
medhelp.org	Health	—	✓
metrolyrics.com	Music, Lyrics	Music & Audio	✓
nytimes.com	News, Newspapers	New - Newspapers	✓
pandora.com	Radio	Music & Audio - Radio	✓
snagajob.com	Employment, Job Search	Jobs & Education	✓
tmz.com	News, Entertainment, Celebrities	Celebrities & Entertainment News	✓
toyota.com	Automotive, Toyota	Autos & Vehicles	✓
tripadvisor.com	Travel	Travel - Hotels & Accommodations	✓
yelp.com	Consumer Opinions	—	—
Miami	N/A	—	✓
skin cancer	N/A	—	✓

Some interest categories listed in Table 7 were shown intermittently. For example, the "Autos & Vehicles" category was often missing, but we constantly received automotive ads.

As an anecdotal example, right after reading a piece of news about a well-known gay person in an uncontrolled browser, we received the Google ad shown in Figure 22 on another site, which included the acronym "LGBT" (lesbian, gay, bisexual, and transgender).

Figure 22: An anecdotal example of ad targeting by Google

4.1.1 Disabling third-party cookies and opting out

During these tests, we did not receive any ads from Google related to our behavior on the sites listed in Table 2. As expected, we received ads relevant to our behavior on the sites listed in Table 3 in the extra sessions.

In the opt-out tests, we observed ads served by non-Google third parties related to our behavior on the sites listed in Table 2, which was also expected. An example of those ads is shown in Figure 23.

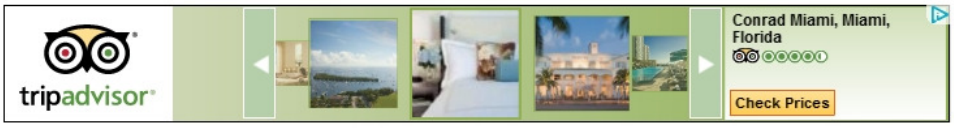

Figure 23: A behavioral ad served by Criteo [8] on `reference.com`

4.2 AOL

Figure 24 shows a snapshot of AOL's APM (the AdVisibility profile) from our tests. Note that AOL does not show demographic information.

Figure 24: A sample snapshot of the AOL AdVisibility profile

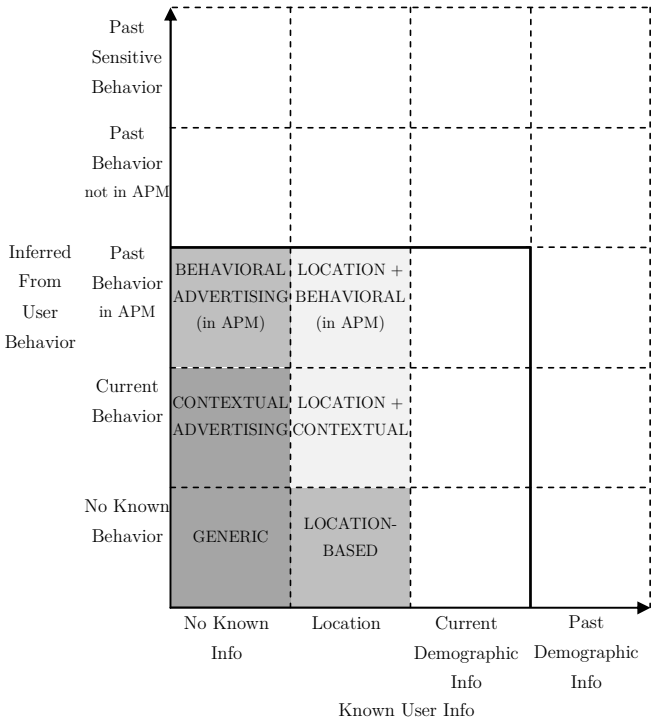

Figure 25: Classification of AOL's advertising practices based on our test results (inner square: expected results, remaining boxes: unexpected results; light gray: observed in some sessions, gray: most sessions, dark gray: all sessions)

We observed AOL ads served by such ad servers as `advertising.com` and `atwola.com`. Figure 25 summarizes AOL's advertising practices we were able to identify. It should be noticed that these are all expected types of advertising.

The AOL test results are summarized in Table 8.

Table 8: AOL test results

Site / Search term	**Category**	**AOL AdVisibility**	**Behavioral ads shown**
`autoblog.com`	Automotive	Automotive	✓
`autos.aol.com`	Automotive	Automotive	✓
`cars.com`	Automotive	Automotive	✓
`encyclopedia.com`	Encyclopedias	—	—
`engadget.com`	News, Technology	Consumer Electronics	✓
`huffingtonpost.com`	News	News & Current Events	—
`latimes.com`	News, Newspapers	News & Current Events	—
`match.com`	Dating	—	—
`music.aol.com`	Music	Entertainment - Music	—
`shoutcast.com`	Radio	Entertainment - Music	—
`slashcontrol.com`†	News, Television († now `huffingtonpost.com/tv`)	Entertainment - Television	—
`spinner.com`	Music	Entertainment - Music	—
`techcrunch.com`	News, Technology	Consumer Electronics	✓
`theboot.com`	Music, Country/Bluegrass Music	Entertainment - Country Music	—
`tourtracker.com`	Music, Tours, Concerts	Entertainment - Music	—
Miami	N/A	—	—

4.3 Yahoo!

Figure 26 presents a snapshot of Yahoo!'s APM (Ad Interest Manager) from our tests. Yahoo! utilizes personal information readily available in the account of the last visitor signed in to Yahoo! using that browser, and states this in its APM.

We observed Yahoo! ads served from such domains as `yieldmanager.net` and `yldmgrimg.net`. Figure 27 summarizes Yahoo!'s advertising practices we were able to identify. Our findings from the Yahoo! tests were similar to the AOL test results—i.e., we observed expected types of advertising.

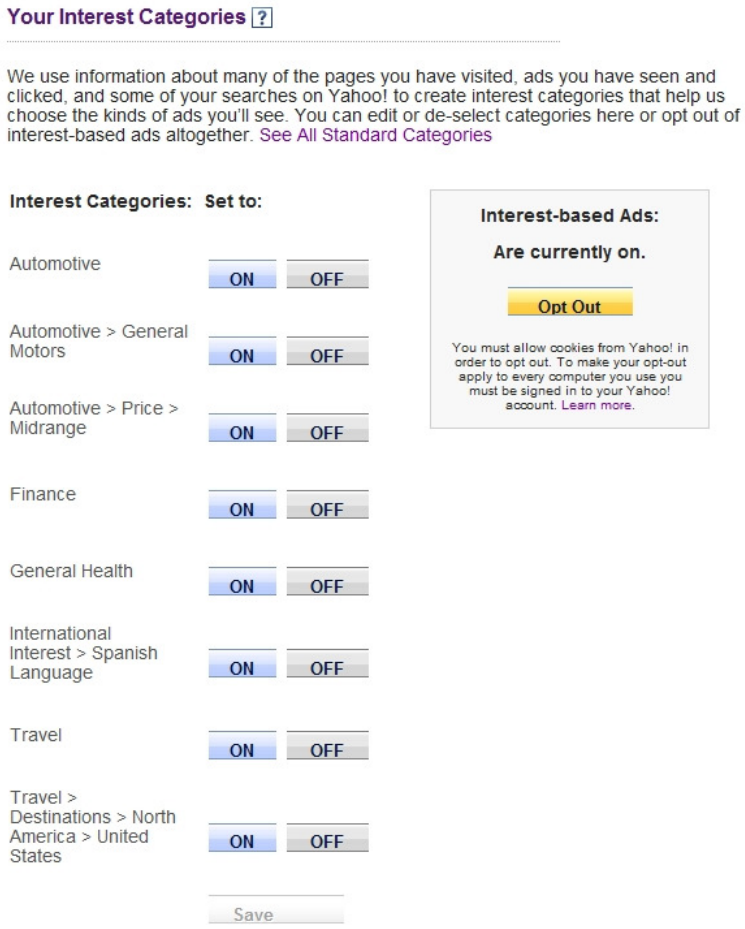

Figure 26: A sample snapshot of Yahoo!'s Ad Interest Manager

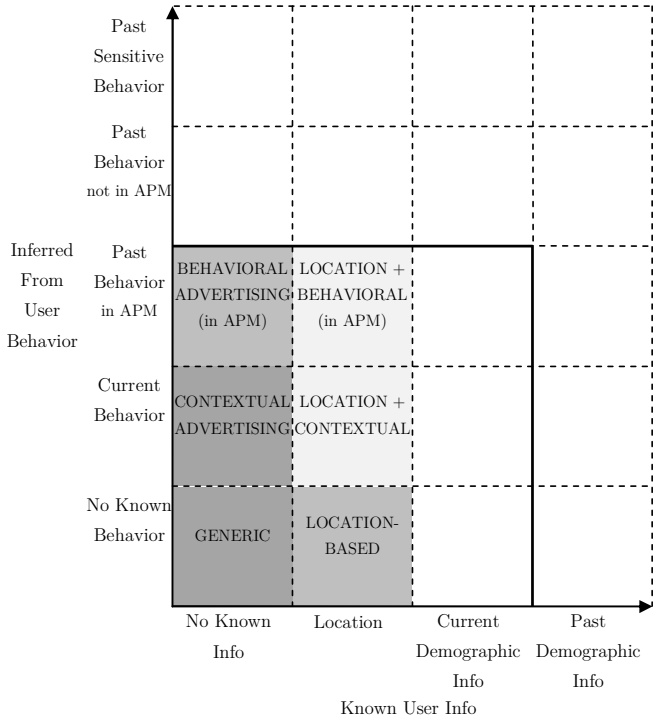

Figure 27: Classification of Yahoo!'s advertising practices based on our test results (inner square: expected results, remaining boxes: unexpected results; light gray: observed in some sessions, gray: most sessions, dark gray: all sessions)

The Yahoo! test results are summarized in Table 9.

Table 9: Yahoo! test results

Site / Search term	**Category**	**Yahoo! Ad Interest Manager**	**Behavioral ads shown**
autos.yahoo.com	Automotive	Automotive	✓
espanol.yahoo.com	Portals, Spanish	Spanish Language	—
health.yahoo.net	Health	General Health	—
hotwire.com	Travel	Travel	✓
lyricsmode.com	Music, Lyrics	—	—
mercurynews.com	News, Newspapers	—	—
monster.com	Employment, Job Search	—	—
nissanusa.com	Automotive, Nissan	Automotive	✓
ozonebilliards.com	Shopping, Sports, Cue Sports	—	—
realestate.yahoo.com	Real Estate	—	—
sfgate.com	News, Newspapers	—	—
travel.yahoo.com	Travel	Travel	✓
webmd.com	Health	General Health	—
wunderground.com	News, Weather	—	—
xfinity.comcast.net	Service Providers, Cable	—	—
Miami	N/A	—	—
skin cancer	N/A	—	—
diabetes	N/A	—	—

4.4 BlueKai

Figures 28, 29, and 30 show snapshots of each section of BlueKai's APM (the BlueKai Registry) from our tests.

As mentioned earlier, BlueKai does not serve ads. Although we were able to identify redirections (HTTP 302) to a number of ad servers, HTTP 200 responses were generally 1x1 GIF images.

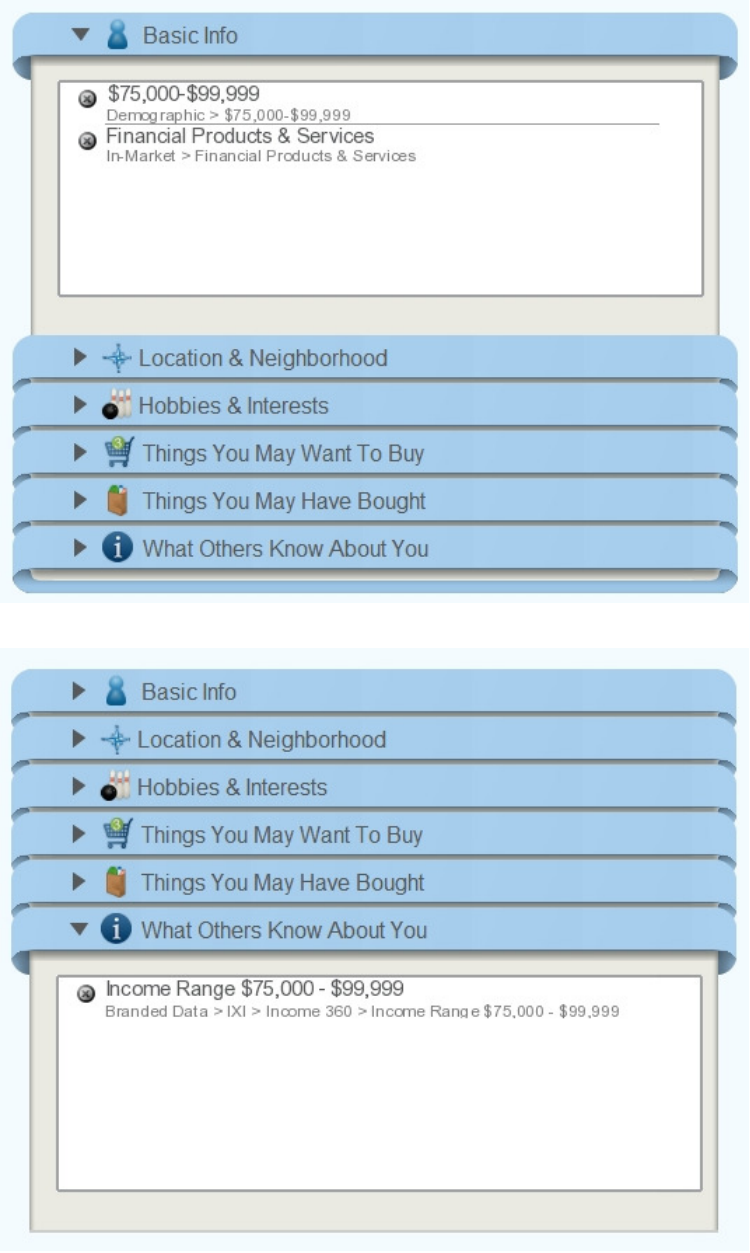

Figure 28: Sample snapshots of the "Basic Info" and "What Others Know About You" sections of the BlueKai Registry

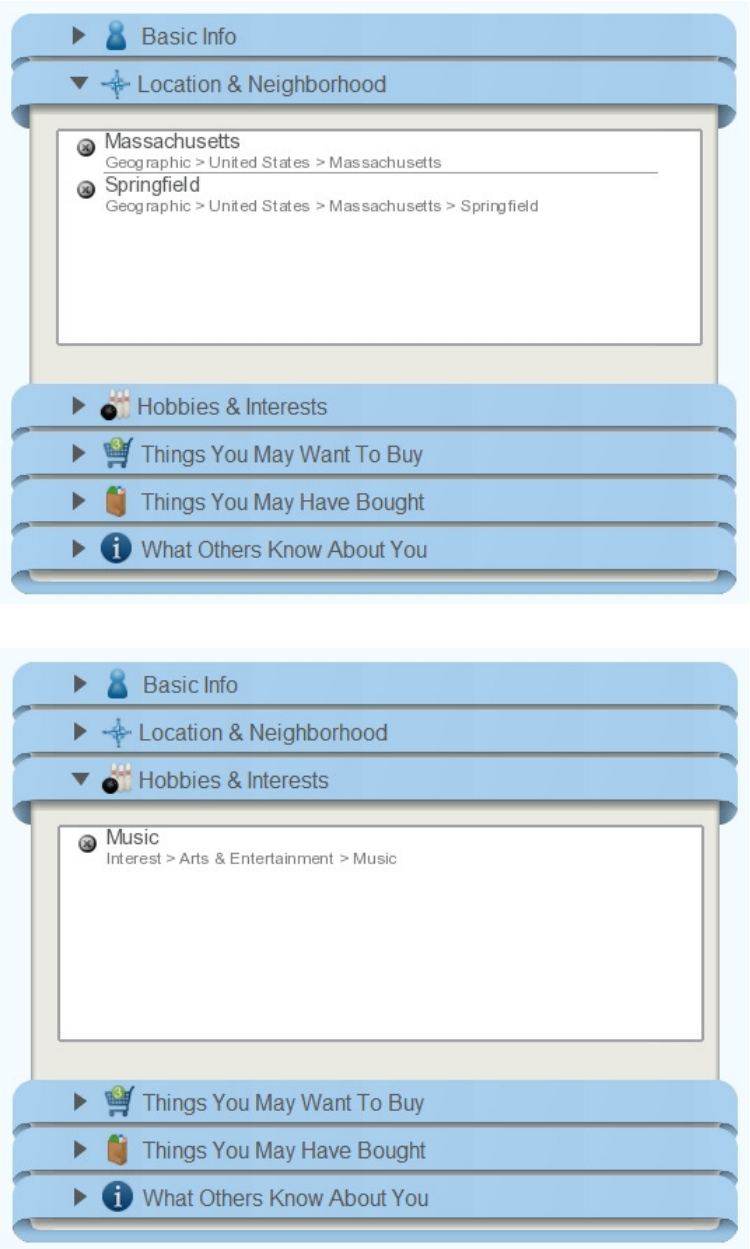

Figure 29: Sample snapshots of the "Location & Neighborhood" and "Hobbies & Interests" sections of the BlueKai Registry

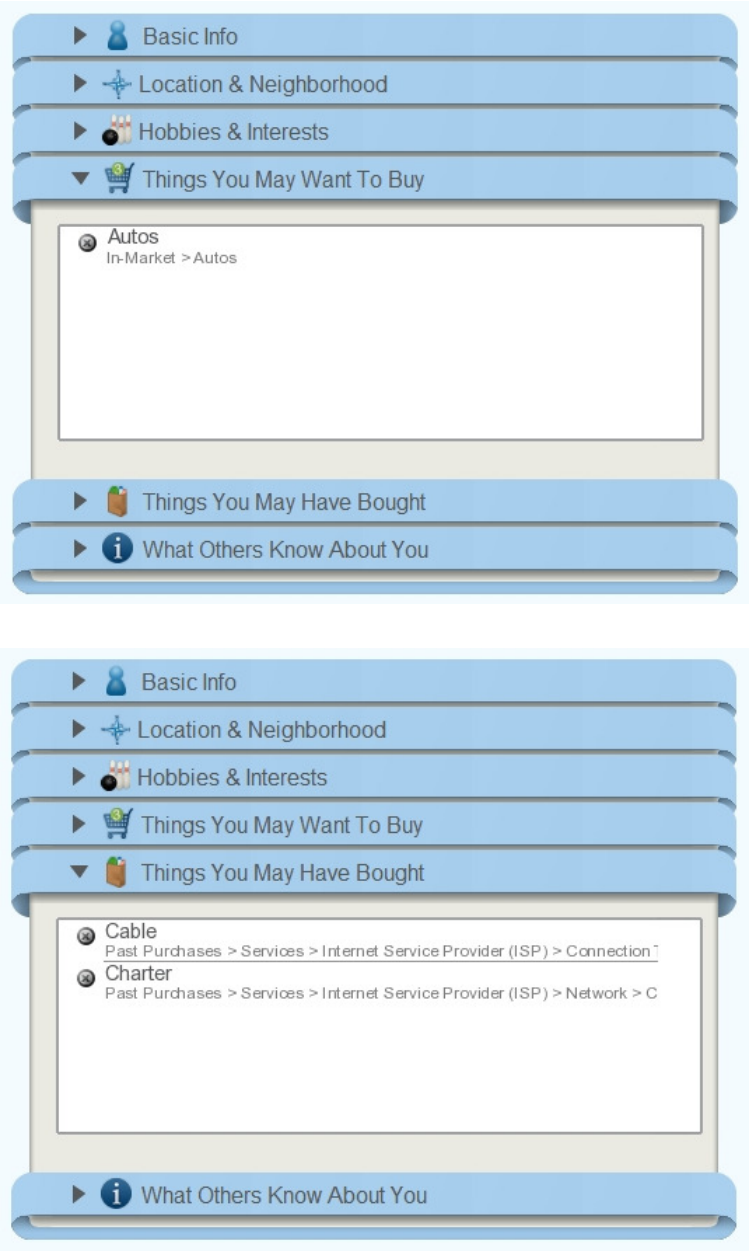

Figure 30: Sample snapshots of the "Things You May Want to Buy" and "Things You May Have Bought" sections of the BlueKai Registry

The BlueKai test results are summarized in Table 10.

Table 10: BlueKai test results

Site / Search term	Category	BlueKai Registry
accuweather.com	News, Weather	—
azlyrics.com	Music, Lyrics	Music
bankrate.com	Personal Finance	Financial Products & Services
bmwusa.com	Automotive, BMW	Autos
cars.com	Automotive	Autos
cbsnews.com	News	—
cdkitchen.com	Cooking	—
cheaptickets.com	Travel	—
contactmusic.com	Entertainment, Music, Movies	Music
gap.com	Shopping, Clothing	—
healthology.com	Health	—
miami.com	Guides, Miami	—
money.cnn.com	News, Business	Financial Products & Services
style.com	Fashion	—
zdnet.com	News, Technology	—
Chicago	N/A	—
New York	N/A	—
skin cancer	N/A	—

4.5 Test with Google properties after March 1, 2012

In this test, we observed the same types of advertising as in Figure 4. We were also able to identify cases in which sensitive information we provided only on Google properties was used to serve ads on non-Google sites. For instance, Figure 31 shows an ad served by DoubleClick on cbsnews.com after searches for "diabetes" on Google.

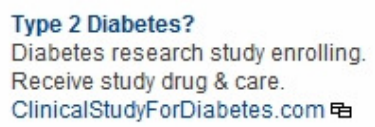

Figure 31: An ad served by DoubleClick on `cbsnews.com` after searches for "diabetes" on Google

We received the ad shown in Figure 32 from another ad service run by Google, Google Ad Services, on `macmillandictionary.com` after watching videos on the Bible on YouTube.

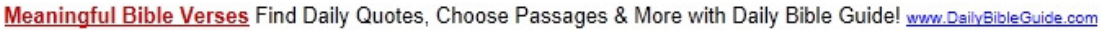

Figure 32: An ad served by Google Ad Services on `macmillandictionary.com` after watching videos on the Bible on YouTube

4.6 Summary

We evaluated the accuracy and completeness of the information in the APMs provided by AOL, BlueKai, Google, and Yahoo!. Another facet of our work was to assess how third parties render ads based on users' browsing behavior. We provided specific instances of advertising practices and illustrated our test results using a two-dimensional classification. Also, we presented sample HTTP headers that demonstrate leakage to the ad server in plaintext. Our findings revealed cases in which even sensitive information was leaked as part of an HTTP header and was used to serve ads on multiple sites.

Chapter 5: Facebook

Although Facebook does not offer an APM, we also examined it in response to the recent debate over its data collection practices [43]. We used a similar methodology for this test to the one explained in Chapter 3. We ran 10 daily sessions with the first-party sites listed in Table 11 and captured all the ads served by Facebook. We never logged out of Facebook and did not "like" any of these sites.

The actions performed on the sites listed in Table 11 included searching for "Chicago" (on `mapquest.com`) and for flights from New York to Chicago (on `expedia.com`).

Table 11: First-party sites used to study Facebook

Site	**Category**
ae.com	Shopping, Clothing
azlyrics.com	Music, Lyrics
bodybuilding.com	Sports, Bodybuilding
cancer.org	Health, Cancer
chevrolet.com	Automotive, Chevrolet
consumerguideauto.howstuffworks.com	Automotive
directv.com	Service Providers, Digital Satellite
drugs.com	Health, Drugs/Medications
entertainment.msn.com	Entertainment
expedia.com	Travel
mapquest.com	Maps
miami.com	Guides, Miami
movies.yahoo.com	Movies
movietickets.com	Shopping, Movie Tickets
wsj.com	News, Business

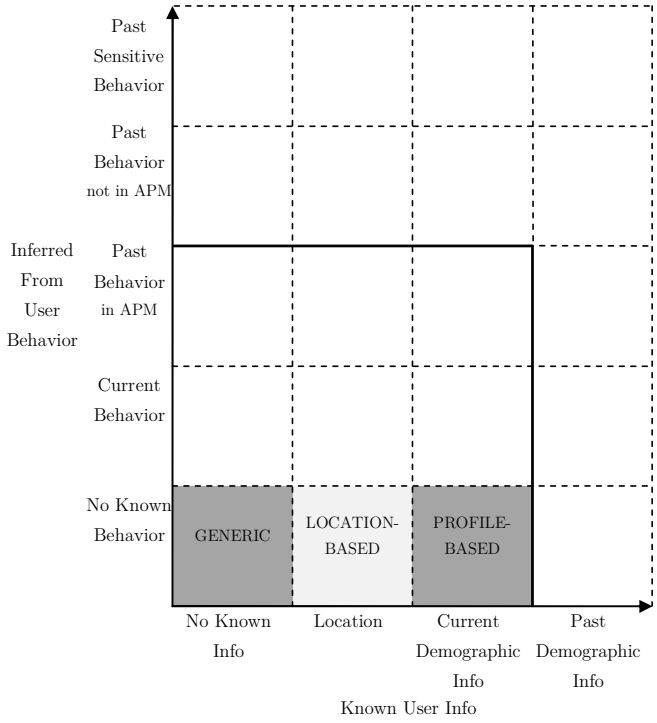

Figure 33: Classification of Facebook's advertising practices based on our test results (inner square: expected results, remaining boxes: unexpected results; light gray: observed in some sessions, dark gray: all sessions)

Figure 33 summarizes Facebook's advertising practices we were able to identify, all of which are expected types of advertising. As an anecdotal example, however, after visiting cancer-related Web sites in an uncontrolled browser, we observed the Facebook ad shown in Figure 34 on a Facebook account not used for testing.

Figure 34: An anecdotal example of ad targeting by Facebook

5.1 Summary

Although Facebook does not offer an APM, we also examined it by virtue of its much-debated data collection practices. For this test, we used a similar methodology to our generic methodology. All of Facebook's advertising practices we identified were expected types of advertising.

Chapter 6: Conclusions and Future Directions

As public concern over behavioral targeting has been growing lately, an increasing number of ad networks offer *ad preferences managers* (APMs) that show collected and/or inferred information about users. In this work, we investigated the accuracy and completeness of the information contained in such APMs and proposed a structured methodology for APM validation. Another facet of our work was to assess how third parties render ads based on users' browsing behavior. We identified expected and unexpected types of advertising and developed a two-dimensional classification to illustrate the advertising practices we observed during our tests. Our findings revealed cases in which even sensitive information was leaked as part of an HTTP header and was used to serve ads on multiple sites.

The rise of computational advertising has triggered systematic research on related systems. There is considerable ongoing effort to develop effective measurement methodologies for online advertising networks. It is important to note

that this study is a snapshot in time. Our methodology can be used to reexamine the third-party aggregators we studied (AOL, BlueKai, Facebook, Google, and Yahoo!) or to examine other third parties. Moreover, it would be possible to gain a better understanding of online advertising practices by automating the testing process.

References

[1] 33Across User Preferences Manager. http://33across.com/preferences.php.

[2] Alexa Internet, Inc. http://www.alexa.com/.

[3] AOL AdVisibility - My Advertising Preferences. http://advertising.aol.com/advisibility.

[4] Bizo - Your Bizo Cookie. http://www.bizo.com/businessProfessionals/your_bizo_cookie.

[5] Bloomberg Businessweek - BlueKai. http://investing.businessweek.com/research/stocks/private/snapshot.asp?privcapId=42917808.

[6] The BlueKai Registry. http://www.bluekai.com/registry/.

[7] Compete, Inc. http://www.compete.com/.

[8] Criteo. http://www.criteo.com/.

[9] S. DeDeo. Pagestats, May 2006. http://www.cs.wpi.edu/~cew/pagestats/.

[10] eXelate Preference Manager for Consumers. http://exelate.com/consumer-opt-out/.

[11] Fiddler Web Debugger. http://www.fiddler2.com/fiddler2/.

[12] J. Freudiger, N. Vratonjic, and J.-P. Hubaux. Towards privacy-friendly online advertising. In *Proceedings of W2SP 2009: Web 2.0 Security and Privacy in conjunction with the IEEE Symposium on Security and Privacy*, Oakland, CA, May 2009.

[13] Google Ads Preferences - Frequently Asked Questions. http://www.google.com/ads/preferences/html/faq.html.

[14] Google Ads Preferences Manager. http://www.google.com/ads/preferences/.

[15] Google Privacy Policy. http://www.google.com/policies/privacy/.

[16] S. Guha, B. Cheng, and P. Francis. Challenges in measuring online advertising systems. In *Proceedings of the 2010 Internet Measurement Conference (IMC)*, Melbourne, Australia, November 2010.

[17] S. Guha, B. Cheng, and P. Francis. Privad: practical privacy in online advertising. In *Proceedings of the 8th Symposium on Networked Systems Design and Implementation (NSDI)*, Boston, MA, March 2011.

[18] M. Helft. Google to offer ads based on interests. *The New York Times*, March 11, 2009. http://www.nytimes.com/2009/03/11/technology/internet/11google.html.

[19] interclick - My Privacy Choice. http://www.interclick.com/privacy/opt-out.

[20] P. G. Kelley, M. Benisch, L. F. Cranor, and N. Sadeh. When are users comfortable sharing locations with advertisers? Technical report CMU-CyLab-10-017, Carnegie Mellon University, 2010.

http://www.cylab.cmu.edu/files/pdfs/tech_reports/CMUCyLab10017.pdf.

[21] D. Kern, M. Harding, O. Storz, N. Davis, and A. Schmidt. Shaping how advertisers see me: user views on implicit and explicit profile capture. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, Florence, Italy, April 2008.

[22] S. Komanduri, R. Shay, G. Norcie, and L. F. Cranor. Adchoices? compliance with online behavioral advertising notice and choice requirements. Technical report CMU-Cylab-11-005, Carnegie Mellon University, 2011.

http://www.cylab.cmu.edu/research/techreports/2011/tr_cylab11005.html.

[23] Korolova. Privacy violations using microtargeted ads: a case study. In *Proceedings of ICDMW '10: The 2010 IEEE International Conference on Data Mining Workshops*, 2010.

[24] B. Krishnamurthy, K. Naryshkin, and C. E. Wills. Privacy leakage vs. protection measures: the growing disconnect. In *Proceedings of the Web 2.0 Security and Privacy Workshop*, pages 1-10, Oakland, CA, May 2011.

[25] B. Krishnamurthy and C. E. Wills. Generating a privacy footprint on the internet. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, pages 65-70, Rio de Janeiro, Brazil, October 2006.

[26] B. Krishnamurthy and C. E. Wills. On the leakage of personally identifiable information via online social networks. In *Proceedings of the Workshop on Online Social Networks in conjunction with ACM SIGCOMM Conference*, pages 7-12, Barcelona, Spain, August 2009.

[27] B. Krishnamurthy and C. E. Wills. Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of the World Wide Web Conference*, pages 541-550, Madrid, Spain, April 2009.

[28] B. Krishnamurthy and C. E. Wills. Privacy diffusion on the web: a longitudinal perspective (updated graphs), October 2009. Submitted as public comment to Federal Trade Commission Exploring Privacy Roundtable Series.
http://www.ftc.gov/os/comments/privacyroundtable/544506-00009.pdf.

[29] Lotame Preference Manager. http://www.lotame.com/preference-manager.

[30] J. R. Mayer and J. C. Mitchell. Third-party web tracking: policy and technology. In *IEEE Symposium on Security and Privacy*, May 2012.

[31] A. Narayanan and V. Shmatikov. Myths and fallacies of "personally identifiable information." *Commun. ACM*, 53, 6 (June 2010), 24-26.

[32] PrivacyChoice - BlueKai. http://www.privacychoice.org/companies/index/20/bluekai.

[33] Quantcast Corporation. http://www.quantcast.com/.

[34] A. Reznichenko, S. Guha, and P. Francis. Auctions in do-not-track compliant internet advertising. In *Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS '11)*, October 2011.

[35] C. Riederer, V. Erramilli, A. Chaintreau, P. Rodriguez, and B. Krishnamurthy. For sale : your data by : you. In *10th ACM Workshop on Hot Topics in Networks (HotNets-X)*, Cambridge, MA, November 2011.

[36] F. Roesner, T. Kohno, and D. Wetherall, Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation*, April 2012.

[37] M. Rotenberg. Google's proposals on internet privacy do not go far enough. *Financial Times*, September 24, 2007.
http://www.ft.com/cms/s/0/764c5338-6a32-11dc-a571-0000779fd2ac.html.

[38] B. Stone-Gross, R. Stevens, A. Zarras, R. Kemmerer, C. Kruegel, and G. Vigna. Understanding fraudulent activities in online ad exchanges. In *Proceedings of IMC '11: The ACM SIGCOMM Conference on Internet Measurement*, 2011.

[39] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 557-570.

[40] TARGUSinfo Profile Manager. http://www.adadvisor.net/profileManager.html.

[41] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas. Adnostic: privacy preserving targeted advertising. In *Proceedings of NDSS '10: The Network and Distributed System Security Conference*, San Diego, CA, March 2010.

[42] J. Turow, J. King, C. J. Hoofnagle, A. Bleakley, and M. Hennessy. Americans reject tailored advertising and three activities that enable it. September 29, 2009. `http://ssrn.com/abstract=1478214`.

[43] J. Valentino-DeVries. Facebook defends getting data from logged-out users. *WSJ Digits*, September 26, 2011. `http://blogs.wsj.com/digits/2011/09/26/facebook-defends-getting-data-from-logged-out-users/`.

[44] Videology Privacy Dashboard. `http://www.videologygroup.com/PrivacyDashboard.aspx`.

[45] Y. Wang, D. Burgener, A. Kuzmanovic, and G. Maci-Fernández. Understanding the network and user-targeting properties of web advertising networks. In *Proceedings of the 31st International Conference on Distributed Computing Systems (ICDCS '11)*, 2011.

[46] C. E. Wills. Identifying and preventing conditions for web privacy leakage. In *Proceedings of the W3C Workshop on Web Tracking and User Privacy*, pages 1-5, Princeton, NJ, April 2011.

[47] Yahoo! Ad Interest Manager BETA. `http://info.yahoo.com/privacy/us/yahoo/opt_out/targeting/details.html`.